Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 69 (2015) 36 – 43

Procedia Computer Science

The 7th International Conference on Advances in Information Technology

Combining Multiple Measures into a Single Figure of Merit

Mark Chignell[a*], Tiffany Tong[a], Sachi Mizobuchi[b], Tamara Delange[a], Wilson Ho[a], William Walmsley[c]

[a]*Univeristy of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8, Canada*
[b]*Vocalage, Canada*
[c]*WhirlScape, Canada*

Abstract

Researchers often collect a number of dependent measures in a study, each of which may reflect some aspect of overall performance. In psychology, in particular response time and accuracy are frequently used measures. Trade-offs between speed and accuracy are often observed, and can occur between other measures. How should overall performance be characterized in the presence of trade-offs? In this paper, we consider how multiple measures can be combined into a single, overall measure of performance. We consider two different case studies, one involving speed and accuracy, and the other involving the relationship between sampling of visual information and resulting accuracy in a pedal-tracking task. We define a global measure of performance using summated z-scores. We discuss the behavior of this measure and contrast it with other approaches to creating linear combinations of variables.

Peer-review under responsibility of the Organizing Committee of IAIT2015

Keywords: human computer interaction; human factors; serious games; speed-accuracy trade-off

1. Introduction

What should researchers do when faced with multiple dependent measures in an experiment? Each variable could be analyzed separately (e.g., with univariate analysis of variance), or the variables could be analyzed jointly (e.g., with multivariate analysis of variance or other multivariate techniques such as discriminant analysis). Discriminant analysis is of particular interest because it builds a discriminant function that maximally discriminates between factor levels in an experiment. When discriminant functions are expressed using standardized coefficients those weights are scaled in terms of standard deviation units. Thus, a discriminant analysis may be conceptualized as a

* Corresponding author. Tel 1-416-978-7581
E-mail address: chignell@mie.utoronto.ca

Peer-review under responsibility of the Organizing Committee of IAIT2015
doi:10.1016/j.procs.2015.10.004

technique for finding a linear combination of z-transformed outcome variables that maximally distinguishes between categories (factor levels) of interest. However, since discriminant functions are fitted to data they will tend to capitalize on error variance, and they will tend to vary considerably from one data set to another. In this paper, we explore a related approach where we construct linear combinations of z-transformed outcome variables without fitting. This approach defines a single evaluative measure that can capture performance across a range of different experiments or contexts. We demonstrate this approach in two different case studies.

2. Background

[1,2] noted a linear relationship (trade-off) between speed and log odds in favour of a correct response. [3] argued that "obtaining an entire speed-accuracy trade-off function provides much greater knowledge concerning information processing dynamics than is obtained by a reaction-time experiment".

[4] cited examples of research where reaction time results may in fact have been artifacts of underlying speed-accuracy trade-offs (SATs). To avoid these problems some researchers have analyzed only correct reaction times. For instance, [5] used response times as the dependent measure after exclusion of high-error participants.

While some researchers were concerned about the implications of the SAT for interpretations of response time results obtained in studies, others examined the SAT as an indicator of neural processes (e.g., [6–8]) and changes due to aging (e.g., [9,10]).

In spite of the relatively large amount of research on SATs in recent decades, there seems to have been relatively little progress towards[2] the original goal of developing a measure (or set of measures) that could characterize both speed AND accuracy in terms of an overall performance measure. However, there has been related work on how to combine mental effort and performance measures (e.g., [11]). In that work distance from standardized z-scores to "a line representing an efficiency of zero" was used to derive an overall measure.

Based on the work of [11],[12] examined different methods of combining speed and accuracy using z-scores. In their case studies, they found that the sensitivity of the resulting global measure depended on the type of relationship that exists in the speed and accuracy measures. In this paper, we extend that work by also examining the use of combined performance measure in a different context (looking at the impact of visual occlusion on accuracy in a pedal tracking task).

3. A Global Measure Of Performance

Fig. 1 shows speed and accuracy in standardized (z-score) coordinates. The zero point at the centre of the coordinates represents mean scores for both speed and accuracy. The y-axis in this figure is defined as a continuous measure of error (e.g., the deviation in pixels between the centre of the target and the centre of the hit point in a tablet-based game). The x-axis is a measure of response time. Due to the way that the axes are defined, a perfect SAT in terms of standardized scores is represented by the line with a correlation of negative one in the figure, while the other line (with positive slope) represents a perfect correlation between the measures (where accuracy gets worse as a longer time is taken to respond).

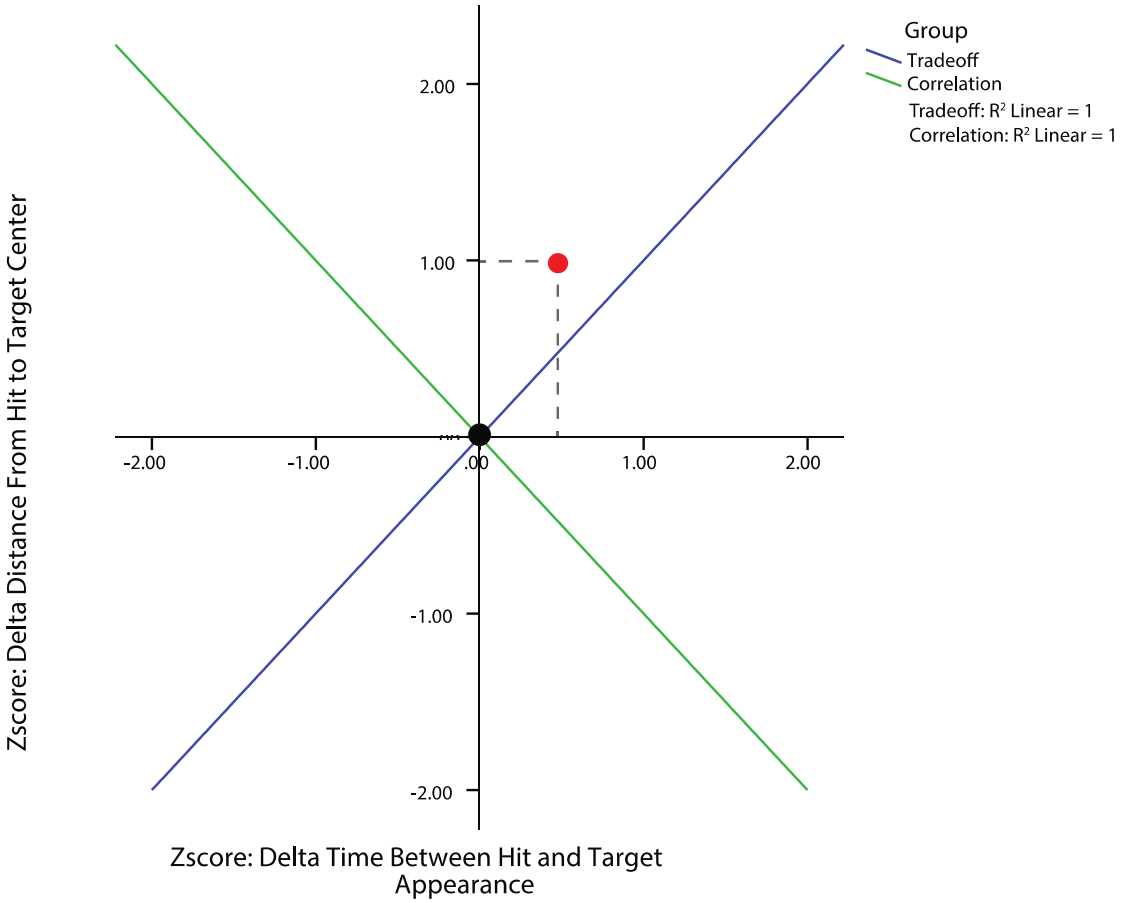

Fig. 1. Performance in standardized coordinates [13].

We can then define a global performance measure as the sum of the distances along each axis from the centre of the axes. In this case, since both axes represent decreasing performance (more error and slower response times) the negated z-scores are summed, i.e., –Z(error) – Z(response time).

This corresponds to a city-block distance from a particular data point to the zero point in the coordinate system (representing mean performance on both the axes). Alternatively, the Euclidean distance metric can be used to define the measure of overall performance by summing the squares of the z-scores of response time and error, and then taking the square root of that value.

It is possible to define a family of measures for a given distance metric (e.g., city block, or Euclidean) by differentially weighting the component measures, e.g.,

A*Z(accuracy) + b*Z(response time)
such that a + b = 1, and
a, b are both contained in the continuous interval [0,1].

Different values of a and b can then be used to give comparatively less or greater weighting to the accuracy or speed components of a global performance measure. Note that while a linear set of weighted variables takes the form of a regression function, we are not advocating fitting the linear function as would occur when using regression analysis or discriminant analysis. Instead, we are suggesting that a set of constant weights by used in the equation, where the weights are either equal (unitary) or else selected, based on a comprehensive set of research findings. In

the case of speed-accuracy, for instance, equal weighting might be appropriate in some domains, but not others. As one example, performance of older participants is often biased towards accuracy at the expense of speed. Thus it may make sense to weight speed and accuracy differently when assessing the performance of different age groups. The advantage of this approach is that it creates a set of standard measurement equations that are not fitted per experiment and that are thus not based, at least partially, on the error variance within a particular experiment.

4. Properties Of The Overall Performance Measure For Response Time And Errors

[13] ran an experiment to assess how well performance on a version of the Whack-A-Mole game could predict the executive function of inhibition ability (e.g., [14]) as assessed by Stroop task performance. The criterion in their study was how well game performance correlated with Stroop task performance.

The response time for a person and a particular combination of game conditions was measured as the median response time in hitting a target (i.e., mole) after it had appeared. The game properties allowed error to be measured as the distance from the user's touch to the centre of the target (in pixels). This relationship was plotted for the entire data set (pooled across participants and conditions) and there was a linear fit ($R^2 = 0.335$), indicating a strong trade-off between median response time and proportion of errors. As users took longer to respond they tended to make fewer errors (i.e., become more accurate). When the data were pooled within participants so that median response time and accuracy were compared across the 24 participants, the correlation between median response time and accuracy (i.e., the between-subject speed-accuracy trade-off) was 0.822. Thus, there was a strong tendency for some participants to be faster than others, but at the expense of accuracy.

Proportion of error and response time scores were standardized across the entire sample by calculating z-scores for the data pooled across participants and conditions. The overall performance score was then calculated as –Z(error)–Z (response time) using the approach outlined earlier. Note that since our interest in this case was in individual differences in ability, it made sense to develop a single distribution for Z(error) scores across the experiment and another single distribution for Z(response time) scores.

Since the goal of the serious game developed by [15] was to measure the executive function of inhibition, the correlation between game performance and Stroop task performance (the Stroop task is a thought to be a measure of inhibition ability) was calculated. No significant correlation was found between Stroop Task performance and game response time, or errors, when considered separately. However, a signification correlation ($r = -0.6$) was observed when the overall performance score, –Z(accuracy)–Z(time), was used (Table 1). This case provides an example of how a combined performance measure may sometimes be more sensitive than individual measures of performance.

Table 1. Correlation between each game performance and median correct response time and percent accuracy on each cognitive ability task. $p < 0.05$, ** $p < 0.01$.

	Stroop Task	Shifting Task	Updating Task
-z(accuracy)	0.011	-0.113	0.220
z(time)	0.257	0.136	0.275
-z(accuracy)-z(time)	-0.600**	-0.399*	-0.354*

The whack-a-mole game was also evaluated with patients from a hospital emergency department. More details on this study can be found in [16]. In this work, performance on the game was compared to standard cognitive assessment scores, where game performance was calculated using the global performance measure (Table 2). In this example, we can see that although –Z(accuracy) and Z(time) have significant correlations with a cognitive test (e.g. DVT), the combined metric, –Z(accuracy) –Z(time) generally had weaker correlations with the cognitive assessment scores than did the standardized response times, Z(time).

Table 2. Correlation between each game performance and performance on standard cognitive assessments. $p < 0.05$, ** $p < 0.01$.

	MMSE	MoCA	CAM	RASS	DSI	DVT
-z(accuracy)	0.173	-0.075	-0.064	0.148	0.155	0.315*
z(time)	0.711**	0.395**	-0.654**	0.340**	-0.734**	0.473**
-z(accuracy)-z(time)	-0.375**	-0.338**	0.412**	-0.099	0.382**	-0.154

A third experiment was then conducted to further assess the effectiveness of the whack-a-mole game in detecting cognitive ability. This study used a sample of 20 able bodied participants varying in age between 20 and 60. Measures of cognitive ability were obtained using established cognitive tests for executive function (the Stroop Task, the Wisconsin Card Sorting Task, and the N-Back Task). Speed (on correct trials) on the cognitive tests were correlated with corresponding performance (speed on correct trials) on the two game variants. For more information on the three cognitive tests see [17]. As can be seen in Table 3, response time was strongly correlated with the cognitive tasks, whereas the combined score was not. Thus in only one of our three examples, comparing performance on a serious game with measures of cognitive ability, did the combined measure of speed and accuracy yield the best results.

Table 3. Correlation between each game performance on the Whacamole task and median correct response time on each cognitive ability task. $p < 0.05$, ** $p < 0.01$.

	Stroop Task	Shifting Task	Updating Task
-z(accuracy)	0.037	-0.119	0.174
z(time)	0.790**	0.598**	0.499**
-z(accuracy)-z(time)	0.037	-0.119	0.174

5. Calculating Overall Performance In An Occluded Tracking Task

[18] conducted an experiment with 16 participants to investigate the effect of cognitive distraction on performance of a 1-d tracking task that simulated gap control in driving. Participants operated the tracking task with a foot pedal while performing a secondary task (an auditorily presented n-back task) under eight conditions involving all possible (2x2x2) combinations of degree of secondary task difficulty (1-back vs. 2-back), pedal tracking difficulty (hard vs. easy amounts of lag) and presence vs. absence of visual occlusion. In conditions involving occlusion, participants could press a button to get rid of the occlusion for a short period of time, i.e., they used voluntary interruption of occlusion (vio).

[18] observed a strong trade-off between the number of times the button was pressed to remove occlusion, and the pedal tracking accuracy (Fig. 2). In this case, accuracy was measured (y-axis) as the target out rate. In order to deal with this trade-off they developed an overall performance measure where number of button presses, and proportion of time outside the pedal-tracking target, were converted to z-scores and then combined. The efficiency of occlusion use was then defined as –Z(Number of button presses) – Z (target out rate). These z-scores were calculated by dividing each condition into eight blocks of equal duration so that there were 32 data means per person (2 levels of pedal tracking difficulty x two levels of N-back difficulty x 8 blocks of time within each condition). The Z-distribution for the two measures (button presses and target out rate) was then calculated across the 32 data means x 16 participants (i.e., 496 data points).

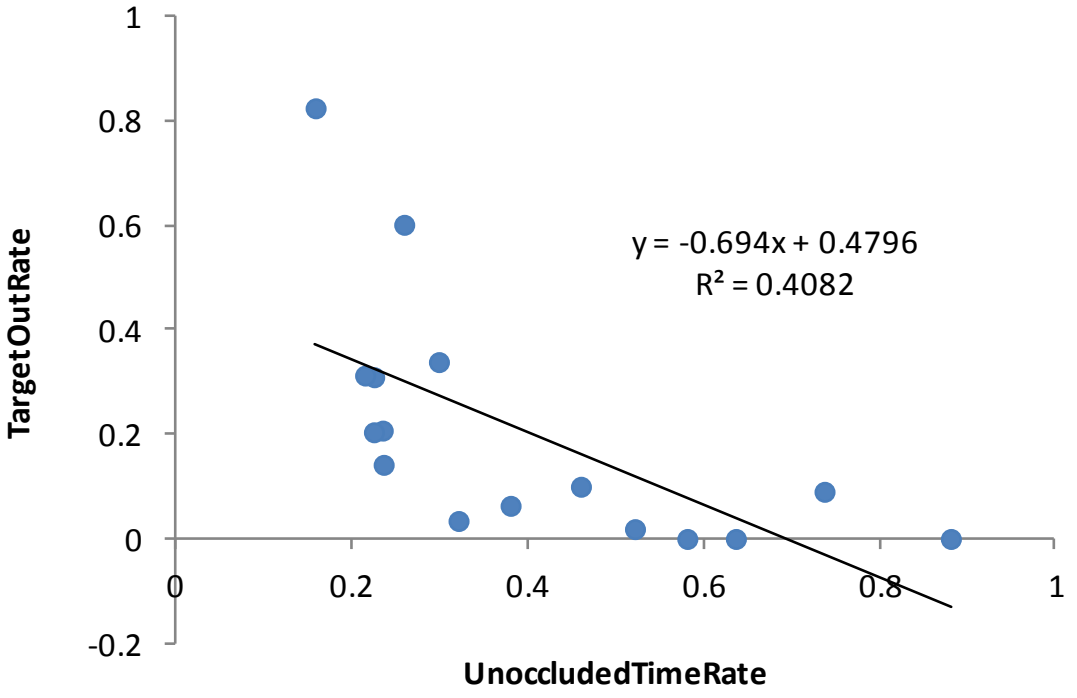

Fig. 2. Trade-off between number of Button Presses and Target Out Rate (each data point is a participant in the study).

The impact of the experimental factors (pedal tracking difficulty and n-back task difficulty) on the overall performance measure (efficiency of occlusion use) in the occluded task condition can be seen in Fig. 3. There is a steady reduction in the overall performance (efficiency of occlusion use) as the difficulty of the primary and secondary tasks increases (note that the bars in the figure are centred around zero because of the use of z-scores). In

this case the efficiency of use measure shows an additive effect of pedal tracking difficulty and n-back (secondary task) difficulty.

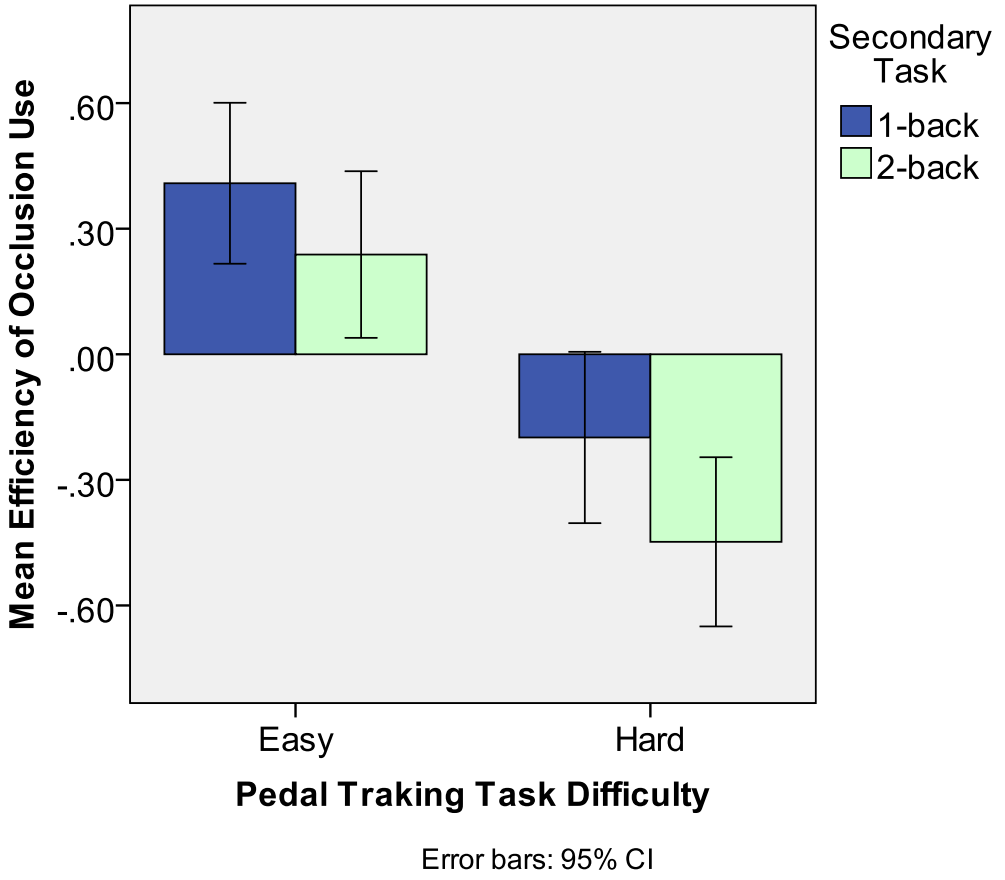

Fig. 3. Success in handling occlusion across levels of primary and secondary task difficulty.

6. Conclusions

Using combined z-score measures to represent overall performance based on a set of separate dependent measures can provide new insights into data and overcome problems in interpreting data due to various trade-offs (such as speed-accuracy trade-offs). While two variable situations are common, there is no reason why the same approach cannot be used with larger numbers of variables.

Our results have shown that combined performance measures using the summed z-score approach works in some cases but not others. While further research needs to be done in order to characterize the situations where these overall performance measures work best, one hypothesis is that they work best when there are moderate (but not too high) correlations between the dependent measures (as is also true for multivariate analysis of variance).

It remains to be seen if overall performance measures are simply a statistical convenience that permits information to be collected from multiple variables at the same time, or if they can measure particular constructs of efficiency or quality of performance in particular contexts. In occluded pedal tracking, for instance, the construct of efficiency of occlusion use or effectiveness of visual sampling makes sense and might apply across a variety of studies. On the other hand, speed-accuracy relationships are highly variable and thus it seems less likely that a particular z-score combination of speed and accuracy will work across a majority of situations.

Our current thinking is that the construction of summated z-score performance measures needs to be done on a case-by-case basis. In some cases, there may be meaningful combined performance measures that represent constructs such as efficiency or quality applying across a range of tasks and situations. In other cases overall measures may be opportunistic. And in some cases, if overall performance measures vary from experiment to experiment it may be simpler to fit them to differences in the conditions for a particular experiment using discriminant analysis.

References

1. Pew, R. W. *Human Perceptual-Motor Performance*. (1974).
2. Pew, R. W. The speed-accuracy operating characteristic. *Acta Psychol. (Amst)*. 16–26 (1969).
3. Wickelgren, W. A. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol. (Amst)*. **41,** 67–85 (1977).
4. Pachella, R. G., Smith, J. K. & Stanovich, K. E. Qualitative error analysis and speeded classification. *Cogn. theory* **3,** 169–198 (1978).
5. Keehner, M., Guerin, S. A., Miller, M. B., Turk, D. J. & Hegarty, M. Modulation of neural activity by angle of rotation during imagined spatial transformations. *Neuroimage* **33,** 391–8 (2006).
6. Osman, A. *et al*. Mechanisms of speed-accuracy tradeoff: evidence from covert motor processes. *Biol. Psychol*. **51,** 173–99 (2000).
7. Heitz, R. P. & Schall, J. D. Neural mechanisms of speed-accuracy tradeoff. *Neuron* **76,** 616–28 (2012).
8. Ho, T. *et al*. The optimality of sensory processing during the speed-accuracy tradeoff. *J. Neurosci*. **32,** 7992–8003 (2012).
9. Salthouse, T. A. The processing-speed theory of adult age differences in cognition. *Psychol. Rev*. **103,** 403–28 (1996).
10. Forstmann, B. U. *et al*. The speed-accuracy tradeoff in the elderly brain: a structural model-based approach. *J. Neurosci*. **31,** 17242–9 (2011).
11. Paas, F. G. W. C. & Van Merrienboer, J. J. G. The Efficiency of Instructional Conditions: An Approach to Combine Mental Effort and Performance Measures. *Hum. Factors J. Hum. Factors Ergon. Soc*. **35,** 737–743 (1993).
12. Chignell, M., Tong, T., Mizobuchi, S. & Walmsley, W. Combining Speed and Accuracy into a Global Measure of Performance. in *Hum. Factors Ergon. Annu. Meet*. (2014).
13. Tong, T. Designing a Game-Based Cognitive Assessment on a Tablet. (2014).
14. Miyake, A. & Friedman, N. P. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Curr. Dir. Psychol. Sci*. **21,** 8–14 (2012).
15. Tong, T. *et al*. Designing Serious Games As Cognitive Assessment Tools For The Elderly. in *Int. Symp. Hum. Factors Ergon. Heal. Care Adv. Cause* **3,** 28–35 (2014).
16. Tong, T. *et al*. Predicting Delirium in Emergency Care With A Tablet-Based Serious. in *Int. Symp. Hum. Factors Ergon. Heal. Care Improv. Outcomes* (2015).
17. Mizobuchi, S., Chignell, M., Suzuki, J., Koga, K. & Nawa, K. The Impact of Central Executive Function Loadings on Driving-Related Performance. in *Adjun. Proc. AutomotiveUI'12* 68–75 (2012).
18. Mizobuchi, S., Chignell, M., Suzuki, J., Koga, K. & Nawa, K. Central executive functions likely mediate the impact of device operation when driving. in *Proc. 3rd Int. Conf. Automot. User Interfaces Interact. Veh. Appl. - AutomotiveUI '11* 129 (ACM Press, 2011). doi:10.1145/2381416.2381437